Syntactically Annotating ASRS Records for Relation Extraction

Tim Miller, William Schuler, Stephen Wu & Lane Schwartz
Department of Computer Science and Engineering, University of Minnesota
{tmill,schuler,swu,lane}@cs.umn.edu

Motivation

- ASRS (Aviation Safety Reporting System) database has free text database of anomalous events.
- The information in this text is very rich, and "bag of words" models miss out on relations in text.
- Relation extraction systems can use syntactic information to recover the relations between referents in text.
- ASRS records present special challenges to syntactic models and require data annotation and a specialized parser.

Goal

Relation extraction using a time-series model parser and domain-specific training and testing data.

Parser Background

This work makes use of a Hierarchical Hidden Markov Model (HHMM)-based parser. This text parser is a specialized version of the parsing engine of a speech recognition system. It makes use of a *right-corner transform* to minimize the amount of memory needed for parsing most text (Schuler et al. 08).

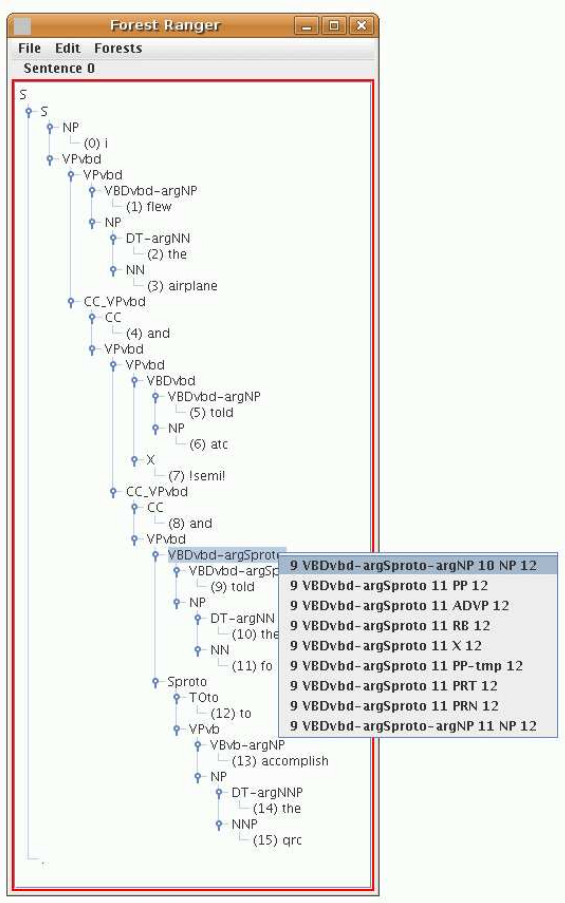
$$\hat{q}_{1..T} = \underset{q_{1..T}}{\operatorname{argmax}} \mathsf{P}(q_{1..T}) \cdot \mathsf{P}(o_{1..T} \mid q_{1..T}) \quad (1)$$

$$\overset{\text{def}}{=} \underset{q_{1..T}}{\operatorname{argmax}} \prod_{t=1}^{T} \mathsf{P}_{\Theta_L}(q_t \mid q_{t-1}) \cdot \mathsf{P}_{\Theta_O}(o_t \mid q_t) \quad (2)$$

In these equations, Θ_L is the language model.
The hidden variable, q, is generalized as a stack of syntactic states, to implement a probabilistic stack-bounded pushdown automaton.

Annotation

- Syntactic annotation involves marking up word strings
- This can be labor intensive and difficult if done completely manually
- Sentences are run through a CKY parser first, so that annotators do not start from scratch.
- In addition, a graphical tool was developed to make annotation easier:



Evaluations on Existing Datasets

Previous work evaluated this system on standard Penn treebank datasets:

- Wall Street Journal: A corpus of newspaper articles – most sentences can be parsed with as few as four stack elements (Schuler et al. 08)
- Switchboard: A corpus of transcribed speech – competitive performance despite complications of speech (Miller and Schuler 08)

Analysis: Memory properties and performance on speech data suggest this model will be successful on the unsegmented and somewhat speech-like text in the ASRS flight records database.

Work Underway

Current Work

- Annotation of sentences from ASRS database (currently about 1000 completed).
- Train parser on a mixture of ASRS and Wall Street Journal data.

Future Steps

- Run experiment to determine inter-annotator agreement.
- Explore techniques to map from syntactic structure to relations
- Build a database of extracted relations